DOCUMENT RESUME

ED 430 039 TM 029 783

AUTHOR Lazarte, Alejandro A.

TITLE Modeling Time To Respond and Probability of Correct Answer in a Simulated Computerized Test-taking Situation.

PUB DATE 1999-04-00

NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.

DESCRIPTORS *Computer Assisted Testing; *Difficulty Level; Higher Education; Probability; *Responses; Simulation; *Test Items; Testing; Time Management; *Timed Tests; Undergraduate Students

ABSTRACT

Two experiments reproduced in a simulated computerized test-taking situation the effect of two of the main determinants in answering an item in a test: the difficulty of the item and the time available to answer it. A model is proposed for the time to respond or abandon an item and for the probability of abandoning it or answering it correctly. In the first experiment, the 47 undergraduate participants were instructed to answer before a certain number of seconds. Results of this experiment suggest that the process to cope with an item is a two-stage decision process in which subjects check to see if the demands of an item match the available time resources. If so, they continue looking for a solution, and if not, they abandon the item. The 48 subjects in the second experiment had to meet an overall deadline of 12 seconds when "tests" or triplets of 3 items were presented. In most subjects, the probability of abandoning the item was still a function of the item's difficulty, but fewer subjects showed an effect of the item's deadline. These results show that under a more realistic test-taking situation, the subject's perception of a deadline to produce an answer is not clearly affected by the external indicators of time availability. This research contributes to the understanding of test-taking behavior in the context of computer-based tests. (Contains 13 figures and 30 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 430 039

Modeling Time to Respond and Probability of Correct Answer in a Simulated Computerized Test-taking Situation

Alejandro A. Lazarte

A. A. Lazarte

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM029783

This paper is prepared for the:
Annual Meeting of the American Educational Research Association in Montreal Canada
April 1999

BEST COPY AVAILABLE

Modeling Time to Respond and Probability of Correct Answer in a Simulated Computerized Test-Taking Situation

Paper presented at the 1999 AERA Meeting - Montreal.

Alejandro A. Lazarte
Department of Psychology
Auburn University
(lazaraa@mail.auburn.edu)

This study reproduced in a simulated computerized test-taking situation the effect of two of the main determinants in answering an item in a test: the difficulty of the item and the time available to answer it. A model is proposed for the time to respond or abandon an item, and for the probability of abandoning or answering correctly the item, which is fitted to data for each subject in the study.

Test-taking is one of the oldest and most common tasks in Psychology, but also one of the most forgotten. Regardless of the psychological attribute to be measured, or the use that we will give to the scores, most of the test-taking situations share the same basic structure: a set of printed questions that the examinees have to answer within a limited amount of time. The observable responses can range from marking a little oval on an answer sheet, to writing a long essay or to putting a puzzle together. The time limit can be so ample that it is not a relevant influence on the outcome, as in power tests, or so short that it is a critical part of the attribute being measured, as in speeded tests. Examinees are given the set of questions and a time limit, and the final recorded information is generally a function of the number of items without answers, the number of correct answered items, and the number of items with incorrect responses. Behaviors the examinees engaged in during the testing situation have been mostly considered in a normative way, i.e., how to behave wisely in a test taking situation in order to improve test scores (Millman, Bishop, & Ebel, 1965).

However, some efforts to generate a theory of test-taking behavior were triggered by recognizing the test-taking task's similarity to many other experimental tasks, where the difficulty of the item, the time restrictions, and the rules for scoring the answers should be combined by the examinees in order to make a decision about answering or not answering a given item.(Baron, Badgio, & Ritov, 1991; Furneaux, 1952; Furneaux, 1961; Pieters, 1985; Thissen, 1983; White, 1973) With the advent of Item Response Theory (IRT) models (Birnbaum, 1968; Lord & Novick, 1968; Rasch, 1960), a renewed interest for integrating cognitive ideas and psychometric modeling was apparent. (Snow & Lohman, 1989). Some models included difficulty components for each item, following a cognitive componential analysis approach (Embretson, 1997; Fischer, 1973; Fischer, 1997; Sternberg, 1985; Whitely, 1977). Other models tried to extend the dimensionality of the latent ability to give a description of complex items (McDonald, 1997; Reckase, 1997). In the same manner, several researchers tackle the relation of the response time in a testing situation with IRT models. Samejima (1973) provided a general two-parameters IRT model for continuous responses that was also applied to response times, and Schebleichner (1979; 1985) extended the Rasch model to response time for corrects answers. These two models assumed "speed" tests with items of a homogeneously low level of difficulty. The items' response times were used as a proxy of the accuracy scores, and, therefore, the relation between response time and accuracy was not explicitly stated.

These models describe the probability of a correct answer as a function of individual differences, item difficulty and time constraints. Some of them also describe the average response time for the item as a function of similar variables. In most of these studies the items were presented one by one, without possibility of reviewing them, and with a continuous recording of time. These models did not have a big influence on applied testing settings because their task did not resemble the actual characteristics of educational or psychological test applications...until now! These test-taking behavior models are bound to gain a new relevance with the arrival of computer-based testing in large commercial applications. The computer transformed the usual paper and pencil test into a task that resembles more a laboratory experimental task, and the continuous recording of response times provided an incredible amount of new information that has recently been used to study the actual test-taking behavior of examinees (Schnipke & Scrams, 1997).

Some conjectures can be made about an individual's behavior in a testing situation with time constraints. For example, subjects read the item and try possible solutions, checking first the simple ones. During these first seconds after the item presentation, the subjects assess the item's difficulty level and the time requirements to obtain an answer. By comparing their estimate of the item's time requirement with the time available for the item, subjects decide to continue working on the item or to stop looking for a solution and abandon the item. We assume that the subject's objective is to accumulate as many points as possible. Under these conditions, the production of a response or an abandonment will

be determined by the combination of the item's difficulty and the comparison of the time requirements with the perceived deadline for the item. These ideas can be expressed as simple logistic regressions where item difficulty and deadline are controlled variables. Following White (1982), we define:

$$Y_{ij} = \begin{cases} 1 & \text{if a subject with ability } a_i \text{ answers correctly an item of difficulty } d_j \text{ and deadline } D_j \\ 0 & \text{Otherwise} \end{cases}$$

$$X_{ij} = \begin{cases} 1 & \text{if a subject with perserverance } r_i \text{ abandons an item of difficulty } d_j \text{ and deadline } D_j \\ 0 & \text{Otherwise} \end{cases}$$

The conditional probability of a correct answer, $C_{ij} = P(Y_{ij} = 1 | X_{ij} = 0)$, and the probability of abandoning an item $A_{ij} = P(X_{ij} = 1)$, are modeled as:

$$\text{logit}(A_{ij}) = \ln\left(\frac{A_{ij}}{1 - A_{ij}}\right) = \alpha_0 + \alpha_1 r_i + \alpha_2 d_j + \alpha_3 D_j \qquad (1)$$

$$\text{logit}(C_{ij}) = \ln\left(\frac{C_{ij}}{1 - C_{ij}}\right) = \beta_0 + \beta_1 a_i + \beta_2 d_j + \beta_3 D_j \qquad (2)$$

where: a_i represents the ability of the individual i ($a_i > 0$), r_i is the persistence in the task of individual i ($r_i > 0$), d_j is the difficulty of item j ($d_j > 0$), and D_j is the deadline for the item j ($D_j > 0$). Thus, for a given individual i, we postulate that the log-odds of abandoning an item and the log-odds for a correct response given a no-abandon are simple additive linear functions of the item's difficulty and the item's deadline. More specifically, we expect that the longer the deadline, the lower the probability of an abandon ($\alpha_3 < 0$) and the higher the probability of a conditional correct ($\beta_3 > 0$). Similarly, the more difficult an item, the higher the probability of an abandon ($\alpha_2 > 0$), and the lower the probability of a conditional correct ($\beta_2 < 0$). Similarly, the higher the accuracy of the individual the higher the probability for a conditional correct ($\beta_1 > 0$), and the higher the persistence, the lower the probability of abandon ($\alpha_1 < 0$).

If we assume that for a given individual i each item is answered independently from the previous ones we can fit the responses of this particular individual simplifying the above models into:

$$\text{logit}(A_j) = \ln\left(\frac{A_j}{1 - A_j}\right) = \alpha_1 + \alpha_2 d_j + \alpha_3 D_j \qquad (3)$$

$$\text{logit}(C_j) = \ln\left(\frac{C_j}{1 - C_j}\right) = \beta_1 + \beta_2 d_j + \beta_3 D_j \qquad (4)$$

where the intercepts, α_1 and β_1 could represent, depending on the scales of the difficulty and deadline variables, the subject's parameters that can be roughly equated to the subject's accuracy and persistence, respectively.

In nearly all tests the time left and the number of items still to be answered at that moment are redefining the task's demands continuously. For example, for the first item, the task starts with n items for t minutes, for the second item, it is redefined as n-1 items for t-t_1 minutes, where t_1 is the time spent on the first item, and so on. This continuous change of the task's time demands will affect the item deadline. A simple indicator of the item deadline that takes into account this sequential redefinition, and the number of items still to answer, is the average time left to allocate per remaining item; for the kth item in a test:

$$M_{(k)} = \frac{t - (t_1 + t_2 + \cdots + t_{k-1})}{n - k + 1}$$

Thus the deadline D_k would be an increasing function of $M_{(k)}$. This characterization of the item deadline produces interesting predictions. For example, for "fast" subjects the test becomes less and less time demanding since the average time per remaining item increases; meanwhile, for "slow" subjects the situation reverses. "Accelerating" subjects are more illustrative of what we might expect in a realistic situation (Schnipke & Scrams, 1997). The subjects start answering slowly reducing the average time per remaining item. However, they start accelerating to catch up with the

overall time limit, and thus, they increase the average remaining time per item. If the time-left and items-left impose a certain degree of demand on the solving of the k item, $k = 1,2,...n$, the performance in the previous k-1 items may also have a similar influence. Let S_k be the accumulated points in the previous k-1 items, then we suggest that the larger the number of accumulated points, the less demanding the task, therefore the longer the subjective deadline. Therefore, a simple expression for the subjective deadline will imply that the greater the average time per remaining time, the longer the deadline per item. Also, the larger the accumulated number of points the more time spent on an item. The simplest function will be a linear one like $D_k = \gamma_1 M_{(k)} + \gamma_2 S_k$. Thus, the models for the log-odds of conditional correct and abandon for an individual i can be expressed as:

$$\text{logit}(A_i) = \ln\left(\frac{A_i}{1-A_i}\right) = \alpha_1 + \alpha_2 d_i + \left(\gamma_3 M_{(i)} + \gamma_4 S_i\right) \qquad (5)$$

$$\text{logit}(C_i) = \ln\left(\frac{C_i}{1-C_i}\right) = \beta_1 + \beta_2 d_i + \left(\gamma_1 M_{(i)} + \gamma_2 S_i\right) \qquad (6)$$

From a review of the literature some empirical regularities with respect to response times in testing situations need to be considered in a holistic conceptualization of test-taking behavior. First, we consider the "regression hypothesis"; i.e. the more difficult an item, the longer the time required to produce a response (Furneaux, 1952). Second, the "filling the time hypothesis"; i.e. when given different time limits, and without affecting their performance significantly, subjects can control their working speed in such a way that the more time allowed, the more time used (Coury & Drury, 1986; Heckman, Tiffin, & Snow, 1967; Pieters, 1985). Following Thissen (1983) and Furneaux (1952) we will assume that the log response times for answered items are normally distributed, with their expected value a function of the difficulty and deadline of the item. We expect that the larger the deadline, the longer the average log time to respond, and the more difficult an item the longer the average log time. For abandoned items, we follow Baron et al. (1991) and postulate that the response time should be only a function of the deadline. For abandoned items we expect that the larger the deadline, the longer the average log time to abandon. Thus, for a given individual i, the log-time for response and abandoned items when the deadline is explicit are given by:

$$\ln(T_A) = u_1 + u_2 D_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma_A^2) \qquad (7)$$

$$\ln(T_R) = v_1 + v_2 d_i + v_3 D_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma_R^2) \qquad (8)$$

The equivalent models in terms of average time left per remaining item and cumulative score will be:

$$\ln(T_A) = u_1 + \left(u_2 M_{(i)} + u_3 S_i\right) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma_A^2) \qquad (9)$$

$$\ln(T_R) = v_1 + v_2 d_i + \left(v_3 M_{(i)} + v_4 S_i\right) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma_R^2) \qquad (10)$$

As we mentioned before, we can speculate that in a testing situation a person changes strategy as a function of a continuous evaluation of the time left and the number of items to go; e.g. if average time for remaining items increases, we will expect the subjects to slow down or keep their pace, but if that average decreases, we will expect faster responses or faster abandons.

The models described above require difficulty and deadline to be controlled experimentally, in comparison to psychometric studies where these parameters have to be estimated. We decided to run two different experiments: one where the available time to work on each individual item, the item's deadline, was explicitly controlled, and another where the time limitations were specified for the whole test.

EXPERIMENT I: ITEM DEADLINE TASK

Method

Stimuli

The stimuli were three-word stem verbal analogies with four alternative choices. A recent and extensive review of results on the analogy items of the GRE provides a criterion to define the analogy's difficulty. Bejar et al. (1991) suggested that the difficulty of a verbal analogy seems to be related in part to the difficulty of the words used, but most importantly to the similarity among alternative response choices: the more similar the alternative responses to the "correct" response, the more difficult the analogy. Of course, "similarity" of alternative responses needs to be defined. An empirical way to operationalize this similarity can be obtained from a report by Ingram & Pellegrino (1977). Additional analogies were obtained from Pintner & Renshaw (1920), that list percent of correct answers to a standard set of 200 analogies. A similar procedure to select verbal analogies was used by Gitomer et al. (1987). The selected analogies were classified in three levels of difficulty using the following criteria: (1) Hard items were considered those analogies where the alternatives included very similar words; (2) Medium items were considered those analogies where two of the alternatives could complete the analogy, while the other two alternatives were totally unrelated to the stem; and (3) Easy items were considered analogies where one-answer was mentioned substantially more often than any other alternative. A set of 48 analogies per level of difficulty were created, giving a total of 144 analogies. These items were assigned at random to 12 lists of 12 items each, with four lists for each level of difficulty.

The item's deadlines were selected in such a way that the minimum deadline would allow the participants enough time to read the stem and all the item options. However, the deadlines were short enough that to abandon the item would be a viable strategy. Some information about average time to answer analogies exists in the literature (Pintner & Renshaw, 1920; Sternberg, 1977; Whitely, 1977). Based on this information, we decided that 3 seconds would be considered as a suitable minimum, in the sense that it would not trigger a decision to abandon the item automatically. We used 5 and 7 seconds as the other deadlines. Some of the items were presented without deadlines to serve as a baseline.

Participants

The participants were 60 unpaid undergraduate students at a Midwest University fulfilling their research participation requirement for an introductory course in Psychology. Of the original 60 subjects, eight subjects produced nearly 90% of all the "fast abandon" responses, i.e. a response time to skip an item smaller than 250 msec. These eight subjects were dropped from the analysis since they did not follow instructions to at least read the analogy. Another five subjects were dropped because they did not abandon any item. Thus, the analyses were performed on an effective sample of 47 subjects.

Procedure

The item deadline task controlled the subjective deadline by instructing the participants to answer before a certain number of seconds, and by utilizing a score system that took into account the time to respond. Incorrect answers were given a flat negative score. If the participants answered before the requested deadline, they received a bonus score proportional to how fast the answer was produced; otherwise they received a penalty proportional to the amount of time their answer went beyond the deadline. Participants still received a bonus when the item was abandoned before the deadline. In similar fashion, a penalty was assigned if the participants abandoned the item after the deadline.

At the beginning of each trial, the deadline for the item was displayed in the center of the screen. The participants controlled the actual presentation of the item. The item appeared in the center of the screen along with a reminder concerning the deadline at the top. At the bottom of the screen a message reminded the subject to press either a "got it" key or a "give up" key as soon as they decide to answer the item or to abandon it. Upon pressing the "got it" key, the screen was cleared and a message asking the participants to key in their answer appeared on the screen. After entering their answers, the correct response, the response time, and the score on the analogy were shown on a feedback screen. The analogy response time was considered the time between the item presentation and the "got it" or "give up" response.

The subjects were instructed to obtain the maximum number of points in each analogy. Also, it was emphasized that to abandon an item was a reasonable strategy, especially when they felt they would not have enough time to answer before the deadline. Guessing was discouraged, and it was stressed to at least read the analogy and the alternatives, even when the item required the shortest deadline (3 seconds). Before the actual 144 experimental trials, the subjects were presented with 12 practice trials. Every subject was exposed to all the combinations of the 3 deadlines plus a no-deadline condition. The order of presentation of the analogies followed a latin square layout with 12 trial blocks x 12 order

positions within each block. The design used the 12 x 12 balanced latin square where each treatment followed and preceded the other treatments the same number of times. The 12 trial blocks were sequenced at random.

Results

Individual Logistic Regressions for Abandons and Conditional Corrects

For Abandons.

The average proportion of abandoned items per subject was 0.24 (St.Dev. = .12), with a range between .06 and 0.61. For each subject, equation (3) was fitted to the responses of the items presented with a specific deadline, a total of 108 items per subject. The fit of the models were assessed using the Hosmer & Lemeshow test of goodness of fit (Hosmer & Lemeshow, 1989) and the chi-square likelihood ratio test for the model. In the case of the goodness of fit test, the average p-value per subject was 0.5930 (St.Dev. = .2390), with a range between 0.0341 and .9761 For only one of the subjects the test did not produce a significant result. For the other forty-six subjects the fit of the model was considered appropriate.

However, the chi-square likelihood ratio test for the model signaled that the model did not increased significatively the likelihood of the data for all the subjects. The p-values for chi-square ranged from less than .0001 to .7783, and for only thirty-seven of the subjects the model was significant (using a critical value of .05). The subjects with a significant model, where categorized in terms of the significance of the parameter estimates for α_2 and α_3. Figure 1 presents the distribution of the subjects according to their fit. Figures 2a to 2c shows the average predicted probability of abandon for each one of the strategy groups:

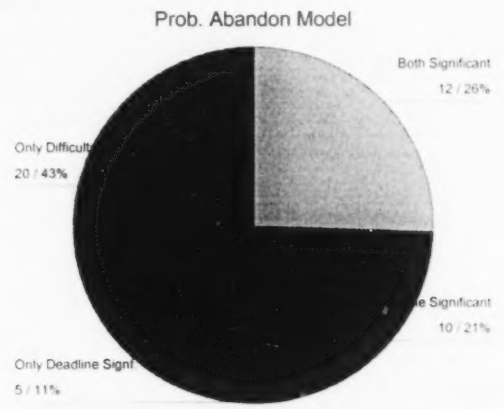

Figure 1:Distribution of fit for the Probability of Abandon Model

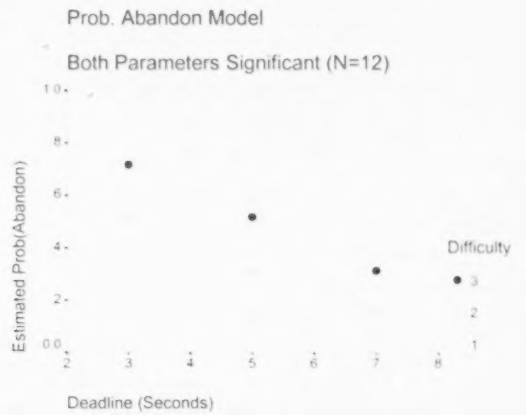

Figure 2a. Average predicted probability of abandon. Both parameters significant

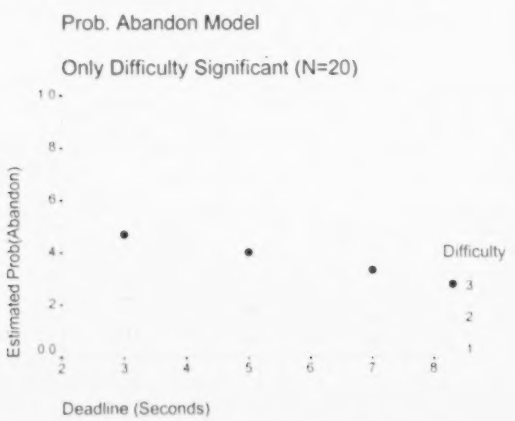

Figure 2b. Average predicted probability of abandon. Only Difficulty significant.

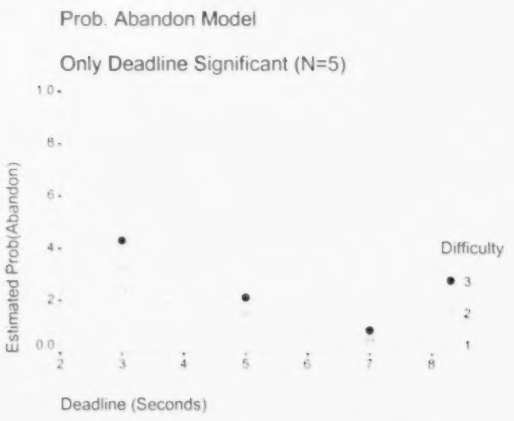

Figure 2c. Average predicted probability of abandon. Only Deadline significant

For Conditional Corrects.

As for the case of abandons, we obtained for each subject in the experiment estimates of the parameters for the model (4) for conditional corrects. For this analysis, the number of observations for each subject varied, since not all of them had the same number of abandoned items. The number of items not abandoned by the subjects varied from 42 to 102 (Mean = 80.3, St.Dev.=13.4). The p-values for the goodness of fit for this model ranged from 0.0285 to .9743 (Mean = .4913, St.Dev.=.3180). The p-values for the chi-square likelihood ratio test for the model with difficulty and deadline ranged from less than .0001 to .0274 (Mean = .0019, Std.Dev. = .0047). In other words, the model was significant, at α = .05, for all subjects in the experiment. Also, the only parameter that was significant was deadline. Figure 3 presents the predicted probability of correct according to level of difficulty.

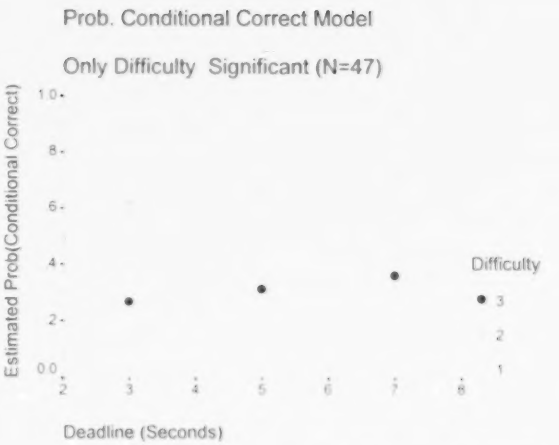

Figure 3. Average predicted probability of corrects. Only Difficulty significant

Analysis of Response Times

Log-Time for Responses. For responses (correct or wrong) we fitted the model (8) to the data of each individual. The average R^2 for this model was .2083 (sd=.0962) with a range between .0279 and .4460 For three subjects the model was not significant, at alpha = .05. For the remaining 44 subjects the average R^2 was .2240 (sd.= .0891). As in the case of the analysis of the probability of abandon or correct response, subjects were classified according to groups with significant parameters. Figure 4 presents the distribution of the subjects according to their fit. Figures 5a to 5c shows the average predicted time for a response for each one of the strategy groups:

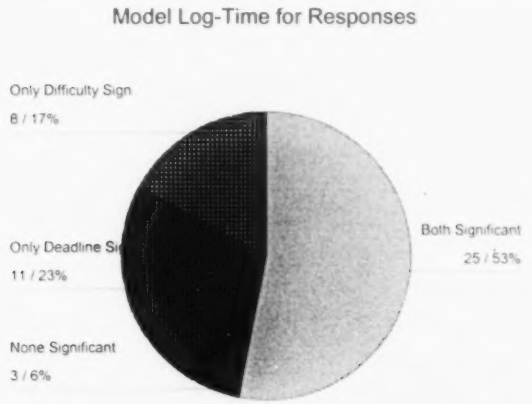

Figure 4. Distribution of fit for the Log-Time of Response Model.

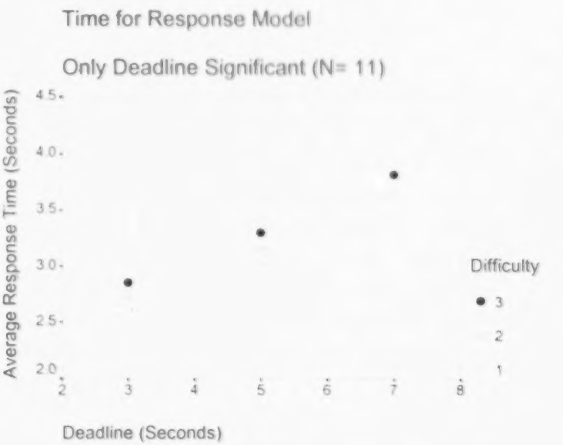

Figure 5c. Average Time for a Response. Only Deadline significant.

BEST COPY AVAILABLE

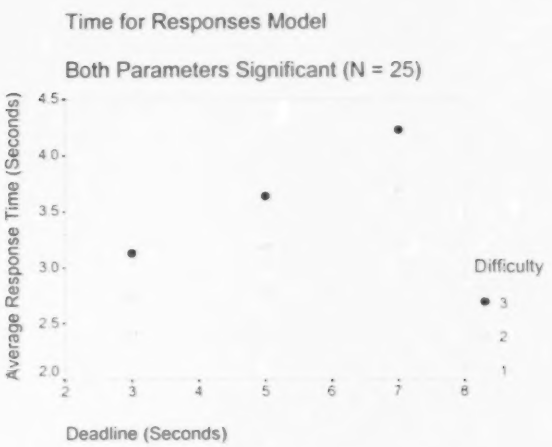

Figure 5a. Average Time for a Response. Both parameters significant

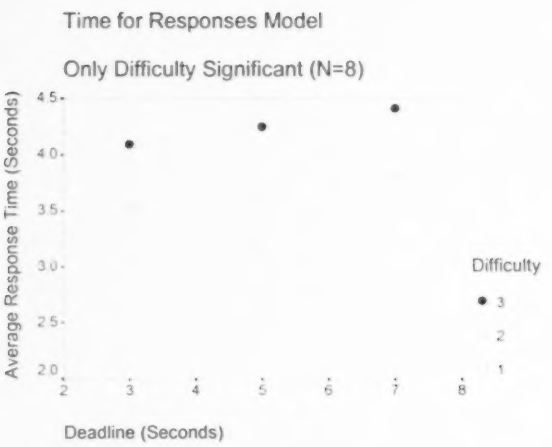

Figure 5a. Average Time for a Response. Only Difficulty significant.

Log Time for Abandons. For abandons we fit the model (7). Even when the model for the time to abandon only includes the effect of deadline, we fit a model with deadline and difficulty in order to check the significance of the difficulty of the item on the log time for abandon. The average R^2 was .2845 (sd=.1766) with a range between .0143 to .7353 In 26 of the subjects the model was not statistically significant ($p > .05$). Subjects in this group seemed to adopt a "flat" time to abandon, in the sense that it was not influenced either by difficulty or by deadline. In the remaining 21 subjects, the only significant effect, as expected, was deadline. For this last group, the average R^2 was .4052 (sd = .1766) with a range between .2194 and .7353 Figure 6 displays the average predicted response time for abandoned items for the 21 subjects with significant deadline effects, and for the 26 subjects with a "flat" strategy, or not significant effect of deadline.

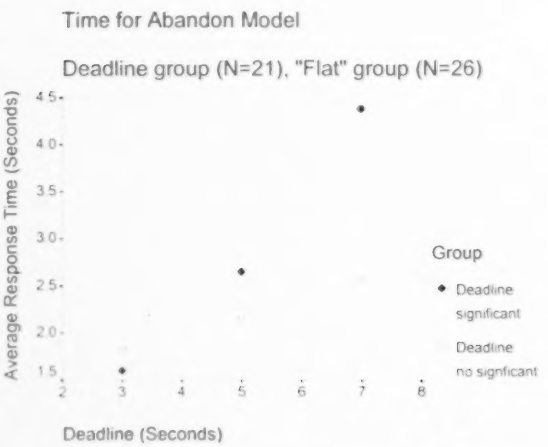

Figure 6. Average response time for abandons, by significance group

Discussion

The results of this first experiment suggest that the process to cope with an item is a two-stage decision process. First, the subjects check if the demands of the item match the available time resources. If this is the case, the subjects continue looking for a solution, otherwise they abandon the item. We suggest that to decide to stop or to continue working, subjects may follow two different strategies: i) to evaluate the task demands while looking for a solution, and to stop when they are reaching the deadline, and ii) to evaluate the task demands as fast as possible and to make the decision to abandon before looking for a solution. The first strategy will explain the influence of a deadline on the average time for an abandon, while the second strategy will explain why item deadlines do not influence time to abandon in some subjects. This also will explain why for nearly half of the sample, the average time for an abandon is shorter than the average time for a response, while for the other half both averages are equivalent. From another perspective, the results of this first experiment follow the patterns of studies involving test-taking under pacing conditions, where subjects adjust their answering speed to the time limitations.

BEST COPY AVAILABLE

EXPERIMENT II: TEST DEADLINE EXPERIMENT

Method

Stimuli

The stimuli were the same 144 three-word stem verbal analogies with four alternative choices used in the first experiment. Forty-eight sets of three items, or *triplets*, were created. Each test contained one easy, one medium, and one hard item. The items within each triplet were ordered in such a way that there were exactly 8 triplets for each of the 6 possible permutations of the three item's difficulties. Based on results from the first experiment, 4 seconds per item was considered as an appropriate deadline per item, therefore the overall deadline for each triplet was 12 seconds.

Subjects

The subjects were 63 unpaid undergraduate students at a Midwest University fulfilling their research participation requirement for an introductory Psychology course. From these, 15 subjects were eliminated because they produced too many quick abandons or not enough abandons, or because they consistently went beyond the overall deadline for the tests. Thus, the analyses were performed on an effective sample of 48 subjects.

Procedure

The Test Deadline task consisted of answering three analogies within an overall deadline of 12 seconds. The actual presentation sequence for each item followed a similar trial layout to that of the Item Deadline Task of the first experiment. At the beginning of each trial, a cover screen displayed the triplet number (from 1 to 48), the item number within the triplet (1 to 3), the accumulated score up to that point and the overall time left before starting the current item. The actual presentation of the item was under the subject's control. The actual display and response procedure was the same as in the first experiment. The subjects were told that the purpose of the experiment was to study how they allocated time to solve problems when facing time limits. The subjects were also instructed to obtain the maximum number of points for each test within the overall time limit. In comparison with the first experiment, the score function for the first and second item in a triplet provided a positive score if the item was answered correctly, and zero points if it was abandoned; no extra points for speed were given. The third item, however, followed the scoring procedure of the Item Deadline task.

Every subject was presented with all the 48 triplets. The six different item orders were organized in blocks of six triplets following a balanced latin square. The blocks defined by the rows of the latin squares were sequenced at random, repeating two of the blocks to obtain the random sequencing of the 48 triplets.

Results

Individual logistic regressions for Abandons and Conditional Corrects

The logistic regression models for the proportion of abandons and conditional corrects, equations (5) and (6) were fitted separately to each subject's data. However, the chi-square likelihood ratio test for the model signaled that for all the subjects the contribution of the accumulated score S was consistently not significant. Thus, we reduced the previous model to a simple version with only item difficulty and time-left per remaining item as independent variables:

For Abandons

The following model was fitted for each one of the subjects:

$$\text{logit}(A_i) = \ln\left(\frac{A_i}{1-A_i}\right) = \alpha_1 + \alpha_2 d_i + \gamma M_{(i)} \qquad (11)$$

This modification did not change the goodness of fit of the model, however the model did not produce a significant chi-square likelihood ratio test not for all the subjects. The p-values for the models ranged from less than .0001 to .7321, and the model was significant for only thirty-one of the subjects (using a critical value of .05). The subjects with a significant model, where categorized in terms of the significance of the parameter estimates for α_2 and γ. Figure 8 presents the distribution of the subjects according to the fit, and figures 9a and 9b shows the average predicted probability of abandon for each group.

BEST COPY AVAILABLE

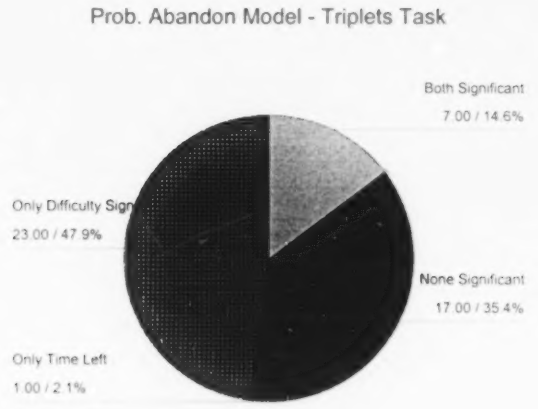

Figure 8. Distribution of fit for the Probability of Abandon Model - Triplet Task

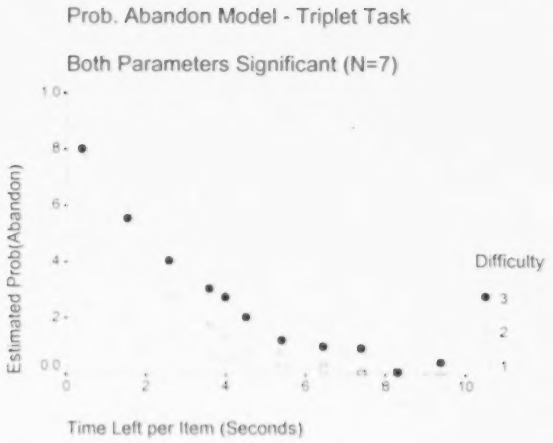

Figure 9a

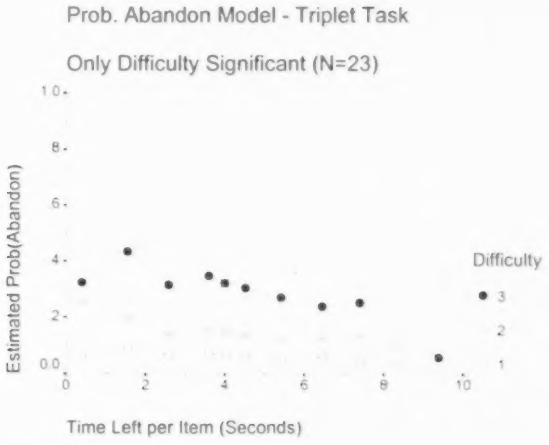

Figure 9b

For Conditional Corrects.

As for the case of abandons, we obtained for each subject in the experiment estimates of the parameters for the modified model for conditional corrects:

$$\text{logit}(C_j) = \ln\left(\frac{C_j}{1-C_j}\right) = \beta_1 + \beta_2 d_j + \gamma M_{(j)} \qquad (12)$$

The p-values for the goodness of fit for this model ranged from 0.0785 to .9743 (Mean = .5213, St.Dev.=.3180). As in the case of the first experiment, all the p-values for the chi-square likelihood ratio test were significant. The p-values ranged from less than .0001 to .0019. However, looking at the individual contributions of difficulty and time-left per item, we found that for two of the subjects these two variables contributed significantly to predict the odds of a correct answer. For the remaining 46 subjects, only item difficulty was significant.

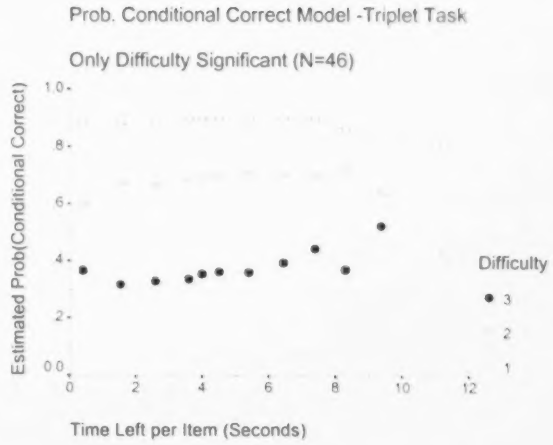

Figure 10b

BEST COPY AVAILABLE

Analysis of Response Times

As with the logistic models, we assumed that for a given subject, the trials were independent and we fitted the models to each individual subject's data. Also, from the original equations (9) and (10), we dropped the cumulative score *S*.

Log-Time for Responses. For responses (correct or wrong) we fitted a model that only includes item difficulty and time-left per remaining item in the triplet:

$$\ln\left(T_{R_{ij}}\right) = v_1 + v_2 d_j + v_3 M_{(j)} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N\left(0, \sigma_R^2\right) \qquad (13)$$

The average R^2 for this model was .1322 (sd=.0785) with a range between .0031 and .3605 The model was not significant at $\alpha = .05$ for seven subjects, that show values of R^2 below .055 For the remaining 41 subjects the average R^2 was .1493 (sd.= .0716). Figure 11 presents the distribution of the subjects according to their fit. Figures 12a and 12b present the estimated average response time for the "Both significant" and "Only Difficulty Significant" groups.

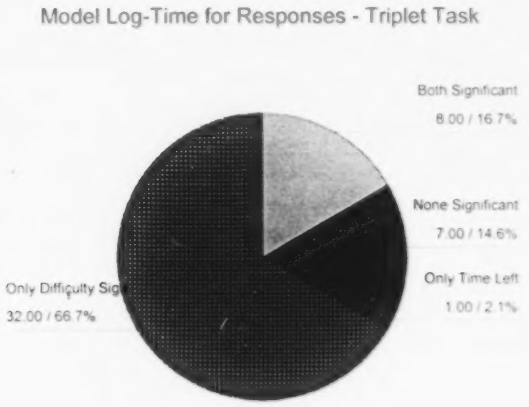

Figure 11. Distribution of fit for the Log-Time of Response Model - Triplet Task

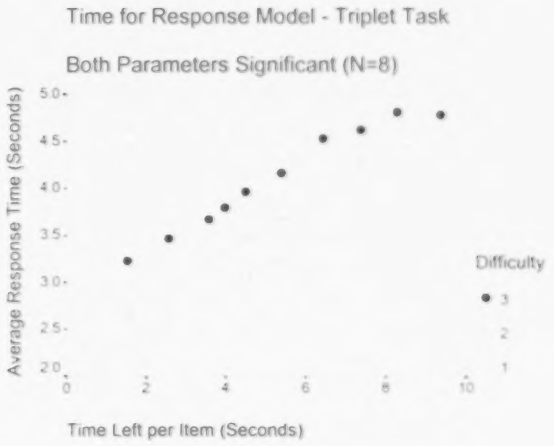

Figure 12a. Average Time for a Response. Both parameters significant Triplet task

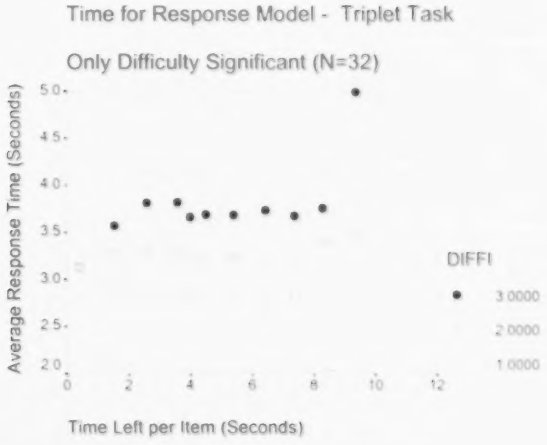

Figure 12b. Average Time for a Response. Only Difficulty significant. Triplet task

BEST COPY AVAILABLE

Log Time for Abandons. For abandons the model (9) was reformulated as:

$$\ln\left(T_{A_i}\right) = u_1 + u_2 M_{(i)} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N\left(0, \sigma^2_{A_i}\right) \qquad (14)$$

where only the deadline level influences the log-time for an abandon. We fit a model that also includes item-difficulty in order to check its contribution. The average R^2 was .2911 (sd=.2398) with a range between .0038 to .8584 In 36 of the subjects the model was not statistically significant ($p > .05$). Subjects in this group seemed to adopt a "flat" time to abandon, in the sense that it was not influenced either by difficulty or by deadline. In 11 of the remaining subjects, the only significant effect was time-left per remaining item. One solitary subject showed a significant effect of item difficulty. For these twelve subjects the average R^2 was .5886 (sd = .1561) with a range between .3302 and .8584 Figure 13 displays this conclusion; the average response time for abandoned items is plotted for the 21 subjects with significant time-left effect. The averages for the group with no significant effect of any variable is overlaid for comparison.

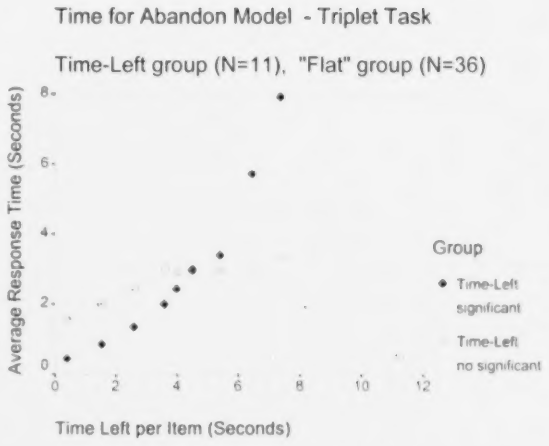

Figure 13. Average response time for abandons, by significance group - Triplet Task

Discussion

In this second experiment, where "tests", or triplets, of three items were presented, and the deadline was given for the whole test, we found a weaker effect of the external measures of time availability, like the average time left per remaining item. In the majority of the subjects, the probability to abandon an item was still a function of the item's difficulty, although fewer subjects showed an effect of the item's deadline. Again, if not abandoned, the probability of a correct response depended only upon the item's difficulty. In comparison with the first experiment, the average time for an abandon did not depend on time left per remaining item in the majority of the subjects. And, the average time for a response (correct or incorrect) was affected by the item's difficulty but not by the time available for the item.

The results of the second experiment show that under a more realistic test-taking situation, the subject's perception of a deadline to produce an answer is not clearly affected by the external indicators of time availability used in this experiment. In fact, the average time for an outcome (response or abandon) remained constant across the three items in each test, even when the time left per remaining item increased. This suggests that the actual deadlines followed by the subjects were constant across the items in the test, leaving item difficulty as the only explanatory variable. This second experiment reflects more the typical situation of a speeded test, where subjects usually work at a constant, and fast, speed.

The advent of computer-based testing as the standard procedure for many commercial tests provides large amounts of new information about test-taking behavior through the recording of response times and answering patterns. In this context, test-taking behavior models are bound to gain new relevance to integrate this new information into a comprehensive theory of a testing situation that includes psychometric as well as cognitive concepts. The present study contributes to the description of test-taking behavior as a function of the two main determinants of an item, its difficulty and the time available to answer it, within a non-psychometric approach.

References

Baron, J., Badgio, P. C., & Ritov, Y. (1991). Departure from optimal stopping in an anagram task. Journal of Mathematical Psychology, 35, 41-63.

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). Cognitive and Psychometric Analysis of Analogical Problem Solving. New York: Springer-Verlag.

Birnbaum, A. (1968). Some latent trait models and their use in inferring examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test score . Reading MA: Addison-Wesley.

BEST COPY AVAILABLE

Coury, B. G., & Drury, C. G. (1986). The effects of pacing on complex decision-making inspection performance. Ergonomics, 29
489-508.

Embretson, S. E. (1997). Multicomponent response models. In W. J. Van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory . New York: Springer-Verlag.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. Acta Psychologica, 37, 359-374.

Fischer, G. H. (1997). Unidimensional linear logistic Rasch models. In W. J. Van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 225-243). New York: Springer-Verlag.

Furneaux, W. D. (1952). Some speed, error, and difficulty relations within a problem solving situation. Nature, 170, 37-38.

Furneaux, W. D. (1961). Intellectual abilities and problem solving behavior. In H. J. Eysenck (Ed.), Handbook of Abnormal Psychology . New York: Basic Books Publisher.

Gitomer, D. H., Curtis, M. E., Glaser, R., & Lensky, D. B. (1987). Processing differences as a function of item difficulty in verbal analogy performance. Journal of Educational Psychology, 79, 212-219.

Heckman, R. W., Tiffin, J., & Snow, R. E. (1967). Effects on controlling item exposure in achievement testing. Education and Psychological Measurment, 27, 113-125.

Hosmer, D. W., & Lemeshow, S. (1989). Applied Logistic Regression. New York: Wiley.

Ingram, A., & Pellegrino, J. W. (1977). Response generation norms for verbal analogies. : Learning Research and Development Center: University of Pittsburgh.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental tests scores. Reading, MA: Addison-Wesley.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. Van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory . New York: Springer-Verlag.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. Educational and Psychological Measurement, 25, 707-726.

Pieters, J. P. M. (1985). Reaction time analysis of simple mental tasks: A general approach. Acta Psychologica, 59, 227-269.

Pintner, R., & Renshaw, S. (1920). A standardization and weighting of two hundred analogies. Journal of Applied Psychology, 4, 263-273.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. Van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory . New York: Springer-Verlag.

Samejima, F. (1973). Homogeneous case of the continuous response level. Psychometrika, 38, 203-219.

Schebleichner, H. (1979). Specifically objective stochastic latency mechanisms. Journal of Mathematical Psychology, 19, 339-350.

Schebleichner, H. (1985). Psychometric models for speed-test construction: The linear exponential model. In S. E. Embretson (Ed.), Test design: Developments in Psychology and Psychometrics. . Orlando FL.: Academic Press.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method for measuring speededness. Journal of Educational Measurement, 34(3), 213 - 232.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Lin (Ed.), Educational measurement (3rd. ed.) . New York: American Council in Measurement and Education.

Sternberg, R. J. (1977). Intelligence, information processing and analogical reasoning: the componential analysis of human ability. Hillsdale, NJ: Lawrence-Erlbaum.

Sternberg, R. J. (1985). Beyond IQ: A triarchic theory of human intelligence. Cambridge: Cambridge University Press.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), New Horizons in testing: Latent trait test theory and computarized adaptive testing. . New York: Academic Press.

White, P. O. (1973). Individual differences in speed, accuracy and persistence: A mathematical model for problem solving. In H. J. Eysenck (Ed.), The measurement of intelligence (pp. 2246-260). Lancaster, England: Medical & Technical Publishing.

Whitely, S. E. (1977). Information-processing on intelligence test items: Some response components. Applied Psychological Measurement., 1, 465-476.